Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies

Journal of proteome research

Article

pubs.acs.org/jpr

Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies

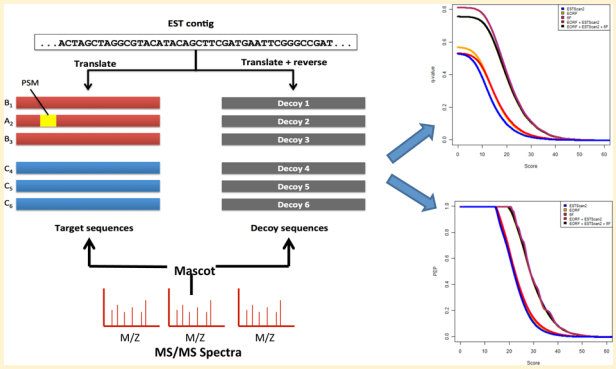
Paul Blakeley,[†,§] Ian M. Overton, [‡] and Simon J. Hubbard*[,†]

[†]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PT, U.K.

[‡]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, U.K.

Supporting Information

ABSTRACT: Proteogenomics has the potential to advance genome annotation through high quality peptide identifications derived from mass spectrometry experiments, which demonstrate a given gene or isoform is expressed and translated at the protein level. This can advance our understanding of genome function, discovering novel genes and gene structure that have not yet been identified or validated. Because of the high-throughput shotgun nature of most proteomics experiments, it is essential to carefully control for false positives and prevent any potential misannotation. A number of statistical procedures to deal with this are in wide use in proteomics, calculating false discovery rate (FDR) and posterior error probability (PEP) values for groups and individual peptide spectrum matches (PSMs). These methods control for multiple testing and exploit decoy databases to estimate statistical significance. Here, we show that database choice has a major effect on these confidence estimates leading to significant differences in the number of PSMs reported. We note that standard target:decoy approaches using six-frame translations of nucleotide sequences, such as assembled transcriptome data, apparently underestimate the confidence assigned to the PSMs. The source of this error stems from the inflated and unusual nature of the six-frame database, where for every target sequence there exists five "incorrect" targets that are unlikely to code for protein. The attendant FDR and PEP estimates lead to fewer accepted PSMs at fixed thresholds, and we show that this effect is a product of the database and statistical modeling and not the search engine. A variety of approaches to limit database size and remove noncoding target sequences are examined and discussed in terms of the altered statistical estimates generated and PSMs reported. These results are of importance to groups carrying out proteogenomics, aiming to maximize the validation and discovery of gene structure in sequenced genomes, while still controlling for false positives.



KEYWORDS: *proteogenomics, peptide spectrum match, false discovery rate, posterior error probability, expressed sequence tag*

INTRODUCTION

Rapid advances in mass spectrometry-based proteomics have been made possible by improvements in peptide separation techniques, high-resolution instruments and downstream informatic processing. Researchers can now obtain a comprehensive catalogue of the proteome in single-celled organisms[1] and a near-comprehensive catalogue in multicellular organisms.[2] However, the database search method that underpins the majority of proteomic workflows typically requires a high quality set of protein-coding genes. To improve proteome coverage, mass spectrometry (MS) data can be searched against protein sequences inferred from either the genome[3−6] or transcriptome.[7−12] Such proteogenomic approaches are not biased toward existing gene annotations and therefore offer scope for novel gene/protein discovery. Indeed, proteogenomics has led to the discovery of thousands of novel gene candidates,[4,5,12] protein isoforms,[13,14] amino acid polymorphisms,[15,16] and confirmation and correction of gene models.[5,17−21] A recurring observation in these studies is that current gene models, which are largely computational predictions themselves, are often incomplete and erroneous. For instance, lowly expressed splice variants and noncanonical genes often prove difficult to annotate.[20,22]

However, the database size and attendant search space when searching raw genomes, particularly metazoan ones, is usually dramatically inflated. For example, a translation of the human genome in all six reading frames would result in a huge search space: typically thousands of LC-MS/MS spectra would need to be searched against at least 6×10^9 amino acids. Moreover, eukaryotic genomes are also heavily populated by non-protein-coding regions, introns and hitherto unannotated splice variants. Although it is possible to search directly against translated genomic sequence, this is clearly a challenging task

Received: May 3, 2012
Published: October 2, 2012

that requires careful quality control to avoid false positives and integration of peptide spectrum matches (PSMs) over a genomic locus.

An alternative approach is to search against expressed sequence tags (ESTs) generated from traditional Sanger sequencing[9,10] or from next-generation sequencing[7,12] since they are by definition transcribed (and likely translated) regions of the genome. For example, this approach has been successfully applied to UniGene clusters by generating a six-frame translation and compressing the resulting protein sequence database to remove redundancy.[9] This allowed the discovery of nonsynonymous mutations, splice-variants, micro exons and alternative translation reading frames, some of which could not be identified from a direct genome search. An alternative method for reducing the search space involves translating the ESTs into proteins using probabilistic approaches such as ESTScan2,[23] DECODER[24] and FrameDP.[25] For example, Robinson and colleagues[10] identified secretory proteins involved in helminth pathogenesis by searching spectra against protein sequences predicted by ESTScan2, which uses a hidden Markov model to distinguish CDS from untranslated regions (UTRs) and correct potential sequencing errors. Such studies have shown that searching against a concise database enriched in protein-coding sequences can increase sensitivity.

Controlling for false positive identifications is essential in high-throughput proteomics studies and particularly so for proteogenomics where intergenic sequence, introns and UTR together can constitute most of the database. Several database search tools such as Mascot,[26] Inspect,[27] X-Tandem[28] or Sequest[29] are commonly used to search spectra against a six-frame translated nucleotide sequence,[4,5,9,17,30−35] but the output scores for candidate PSMs cannot be compared directly across experiments. Indeed, it is widely accepted that applying *E*-value thresholds is anticonservative[36] and different search engines estimate quite different significance levels.[37] Instead, the global error rate is typically estimated by ordering the PSMs according to their *E*-value or search engine-specific score and calculating the false discovery rate (FDR).[38−40] The FDR is estimated from the percentage of incorrect PSMs at a given threshold, which is usually calculated using a target-decoy approach. The exact details of this vary and much discussion in the field exists as to the best approach,[41−45] although the basic principle is well established. This involves searching the spectra against both a target database of 'real' sequences and a decoy database containing 'fake' sequences produced by reversing or randomizing the target sequences; hits to the latter are considered to be false and are used to estimate the level of incorrect target PSMs at a given score threshold. The FDR applies globally to a collection of PSMs, but individual PSMs can also be associated with a *q*-value[38] from the FDR level at which they are first reported. The *q*-value is still a measure of the global error rate within a set of PSMs and like the FDR is dependent on the properties of the database used.[36]

In contrast to the global error rate, a 'local' FDR termed the Posterior Error Probability (PEP) is also frequently estimated as the probability of an individual PSM being incorrect. For example, the software tool *Qvality*[46,47] implements a non-parametric approach for calculating the PEP by estimating the proportion of the target score distribution that is incorrect given a set of *p*-values or a decoy score distribution. Similarly, PeptideProphet[48] can be used to calculate PEPs and can incorporate decoy search results to enable semiparametric modeling and therefore greater flexibility.[49]

Both the FDR and PEP approaches exploit target/decoy database search results to assign statistical significance to PSMs, presuming that the target database accurately represents genuine protein sequences and the decoy database is of equal (or known) size and similar redundancy.[44] In this study, we examine whether these criteria are indeed met in the context of proteogenomics experiments. Specifically, we highlight the problems associated with the standard target−decoy approach for assessing the statistical significance of PSMs assigned to predicted protein sequences derived from a large collection of Chicken ESTs. We consider searches against six-frame translations and single-frame predicted protein translations, comparing different approaches to estimate the statistical significance of the PSMs. Search results are highly dependent on database choice and suggest potential pitfalls when searching six-frame databases linked to database size and target−decoy error modeling. We believe this leads to overconservative significance estimates for six-frame translation databases and a reduction in sensitivity and, hence, fewer confident PSMs and peptides. We investigate the source of the errors, biases in six-frame translation databases, and suggest a variety of approaches to address these problems. This can be achieved by modifying the six-frame database or by conceptually translating the EST sequences. The results have significance for any group carrying out proteogenomic searches against genomic or transcriptome-based sequences.

MATERIALS AND METHODS

EST Data Set

A total of 339 314 ESTs were sequenced from 64 cDNA libraries derived from 21 chicken tissues, and then clustered and assembled using BLASTN and PHRAP to generate 85 486 EST contigs.[50] These are available via http://www.chick.umist.ac.uk. Two different sets of protein sequences were predicted from each EST contig using the EORF and ESTScan2[23] algorithms.

Preparation of Chicken Samples and Mass Spectrometry

We used a comprehensive data set of peptide spectra generated for an unrelated DT40 project, kindly donated by colleagues at the University of Cambridge (Kathryn Lilley, personal communication). The MS/MS data were derived from a proteomic analysis of the DT40 chicken cell line which used the LOPIT protocol, and was originally published in 2009.[51] Full details are available in the original paper,[51] but briefly, DT40 cells were fractionated by density gradient centrifugation and 7 fractions were chosen for analysis. The selected fractions were labeled with four-plex iTRAQ reagents and digested with trypsin. Labeled peptides were pooled together and separated using two-dimensional liquid chromatography. LC-MS/MS was performed using an ultimate-nano-LC system (Dionex) coupled to a QSTAR XL mass spectrometer (Applied Biosystems). The QSTAR XL was operated in information-dependent acquisition mode in which 1 s MS scans were performed (400−1600 m/z) followed by 3 s product ion scans (100−1580 m/z) on the two most intense doubly and triply charged peptides. The LOPIT protocol and iTRAQ labeling were incidental to our study which focuses solely on the relative merits of database composition and attendant statistical treatments to peptide identification.

dx.doi.org/10.1021/pr300411q | J. Proteome Res. 2012, 11, 5221−5234

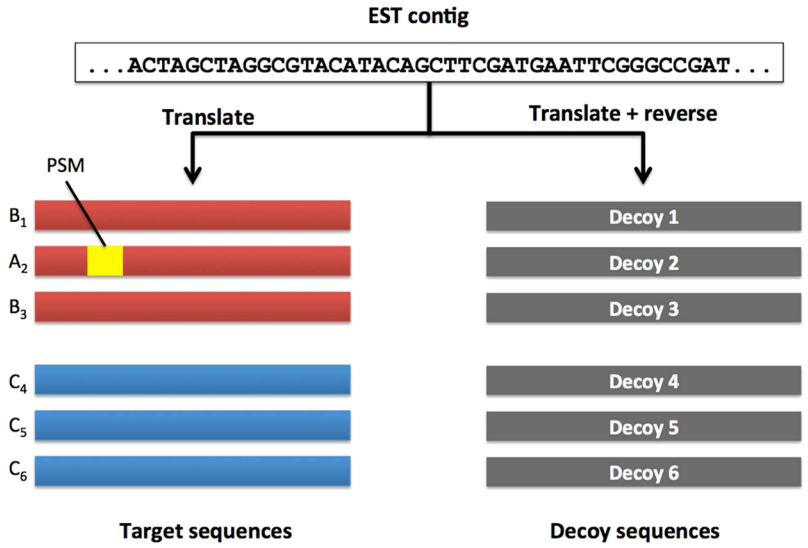


Figure 1. Schematic of EST translation for target:decoy database generation. Translation of transcriptome data such as ESTs in all six reading frames increases the proportion of 'junk' sequence. In this simplified model, only one of the six reading frames is correct (sequence A in frame 2). Sequences denoted by "B" are in the correct direction and therefore in some circumstances could constitute part of the correct ORF as a result of pre-mRNA splicing or frameshift errors. Sequences denoted by "C" are in the wrong direction and are therefore incorrect. Decoy sequences are created by reversing the six corresponding target six sequences, so that decoy1 is the reverse of B_1, decoy 2 the reverse of A_2, and so on.

Databases

Ensembl protein sequences were downloaded in fasta format from the Ensembl FTP server (ftp://ftp.Ensembl.org). Version 56 of the 'pep.all' set containing translations from known and novel genes was used. The UniRef90 database (release 15.11) was downloaded from (http://www.uniprot.org/downloads). A custom Perl script was used to generate the six-frame translation database from 85 486 EST contig sequences, generating 3 forward and 3 reverse-frame sequences using the standard genetic code. Protein sequences were also predicted from the EST contigs using the ESTScan2 and EORF programs (manuscript in preparation). Briefly, EORF calculates a score for each reading frame based on its codon usage bias and sequence homology derived from BLAST searches againt UniProt, and was used to predict 67 125 ORFs from the EST contigs (a small fraction are rejected as unlikely to be coding). ESTScan2[23] uses a hidden Markov model to predict the correct reading frame and UTRs. ESTScan2 predicted a total of 62 161 protein sequences from the ESTs (again rejecting noncoding ESTs).

Mass Spectrometry Database Searching

A total of 403 820 centroided spectra were searched against protein sequences derived from the ESTs using Mascot version 2.0, with a precursor MS error tolerance of 0.2 Da and MSMS error tolerance of 0.8 Da. Fixed modifications were Cysteine β-methylthiolation and iTRAQ labeling of Lysine residues and N-termini. Variable modifications were iTRAQ labeled tyrosine and Methionine oxidation. Up to 1 missed cleavage was permitted.

Five different databases were searched: Ensembl version 56, EORF predictions, ESTScan2 predictions, and the six-frame translations. In addition, various combinations of EORF, ESTScan2 and six-frame sequences were searched to find the database which allows for the most PSMs. Typically, the majority of PSMs for a database search are incorrect; hence, it is important to accurately predict which PSMs represent real peptides present in the samples. For statistical evaluation of the data, decoy databases were constructed by reversing each protein sequence in the original 'target' database. Two types of target−decoy searches were performed: separate and composite. In the composite search, the decoy database was concatenated onto the target database and then used as a single database for the Mascot searches, whereas the separate search involved independent searches against target and decoy databases.

Statistical Methods for Validating PSMs

From composite (concatenated) searches, FDRs were calculated using the methods published by Ellias and Gygi,[40] and Käll et al.[39] to estimate the proportion of false positive PSMs that have accumulated at a given Mascot score giving FDR_{EG} and $FDR_{Käll}$, respectively. Because the FDR does not increase monotonically with Mascot score, q-values were calculated as the minimum FDR at which a PSM is accepted. The *FDRScore* method[52] was used to combine PSMs from the EORF and ESTScan2 searches. For each of the two database searches, a custom perl script was used to assign q-values against Mascot scores, to identify step points (where the q-value increases). From this, a linear regression was calculated between each step point. The *FDRScore* was calculated for each Mascot score between the step points, according to the gradient of the line. PSMs common to both searches were then merged by calculating the geometric mean of their *FDRScores*. The PSMs were then resorted by their Average *FDRScores* to calculate a new set of q-values, from which a second *FDRScore* was calculated, termed the *combined FDRScore*.

In parallel to the composite database estimates, the Mascot scores for separate database searches were used to calculate the local FDR, or PEP, using the software tool Qvality.[46,47] Default parameters were used to generate a set of PEPs and PEP-derived q-values linked to Mascot scores.

Estimating the Proportion of Correct PSMs

EST contigs and their attendant reading frames were assigned to Ensembl proteins via BLASTX searches against the Ensembl56 database. Assignments were made for the top

dx.doi.org/10.1021/pr300411q | J. Proteome Res. 2012, 11, 5221−5234

scoring hits which passed the following cut-offs: Identity >95%; Coverage >50 residues; *E*-value <0.001. EST contigs with significant hits were extracted along with the top scoring reading frame. We assumed that the highest scoring reading-frame contained the correct ORF. This information was integrated with the six-frame translation database, to identify the sequences in the 'correct' reading frame owing to the BLASTX match. PSMs having a match with the correct reading frame were then assumed to be correct.

The probability distributions of the amino acid frequencies were calculated for each reading frame and compared with the probability distributions for the Ensembl and UniProt90 protein sequences. The amino acid frequency of the entire correct reading frame set was calculated separately from the incorrect reading frames. The Mann−Whitney U test was used to measure the degree of divergence between the amino acid distributions of the correct frames with each of the incorrect frames different distributions.

RESULTS AND DISCUSSION

Searching against Six-Frame or Redundant Databases Affects Sensitivity

When searching high-throughput mass spectrometry data against a protein database using a target−decoy strategy, it is usually the case that the target database is composed of genuine protein sequences that could be present in the sample, and random (false) matches to target and decoy database are equally likely. However, when searching against a six-frame database, these conditions are not necessarily met. One anomaly is that at most only one out of the six possible reading frames translated from a nucleotide sequence is likely to be coding (presuming that there is only one protein coding ORF at any given locus) and can lead to "true" target PSMs. This is illustrated in Figure 1 for a hypothetical EST sequence that codes for a protein, along with its six conceptual translations (forward frames 1, 2, 3 and reverse frames 4, 5, 6) and the attendant six decoy frames. Only the target frame A2 contains a true coding sequence and is matched by a single PSM in this case. Frame B1 and B3 are in the correct direction (and could contain correct PSMs if there were a frameshift mutation in the nucleotide sequence) while frames C4−C6 are not. Although the target and decoy databases contain the same number of sequences, amino acid composition and tryptic peptides, the five "wrong" target frames are not likely to be protein-like. This expansion of the target database is in principle similar to simply adding more proteins (perhaps from another species) to the database, but in this case, the additional targets are clearly "wrong" and not likely to be protein-like in composition. Moreover, the expansion is particularly big, adding five extra sequences for every original target. This could confound the assumptions normally held for target-decoy FDR calculations and lead to incorrect statistical modeling.[45]

To test this, we searched tandem mass spectra derived from a chicken DT40 cell line against different protein databases generated from assembled EST contigs, applying a variety of FDR and PEP-based confidence measures to generate significant PSMs. For FDR estimation, two widely used approaches using concatenated target−decoy databases were used based on methods published by Käll et al.[39] and Elias and Gygi[40] to generate $FDR_{Käll}$ and FDR_{EG} estimates with attendant *q*-values. Finally, the PEP and PEP-derived *q*-values were also estimated from separate database searches using Qvality.[47] As described in the methods, five different protein sequences databases were generated from the EST contigs and assessed by the number of PSMs accepted at a consistent threshold (*q*-value/PEP < 0.01) for the different confidence measures (Table 1). Decoy sequences were created by reversing the target sequences (see also Materials and Methods).

Table 1. Unique Peptide Identifications at *q*-Value/PEP < 0.01 for Different Database Searches

	FDR_{EG} *q*-value	$FDR_{Käll}$ *q*-value	PEP-derived *q*-value	qvality PEP
ESTScan2	10519	12291	10721	7323
EORF	9606	11502	9290	7143
six-frame translations	6730	8620	7328	4951
EORF + ESTScan2	9702	11466	9854	6962
EORF + ESTScan2 + six-frame	6778	8616	7405	5020
EORF + ESTScan2 (*FDRScore*)	11532	13813	-	-

Table 1 shows that considerably fewer PSMs were accepted when searching the six-frame database at a *q*-value/PEP cutoff of 0.01, in comparison to the other databases searched. This observation is independent of the chosen measure of statistical significance, suggesting it is a problem specific to the six-frame database searches (whether concatenated or separate). This reduced sensitivity can be attributed to several factors. First, the six-frame database is more than seven times greater in size than the other databases, and it is known that large databases can lead to reduced search sensitivity owing to more conservative statistical estimates[45,53] (or conversely, small database size leads to overestimates of significance and likely false positives). Second, the unusual nature of the six-frame database could be confounding and inflating the FDR. In all likelihood, only one of the six forward frames is "correct" and can be matched by genuine peptide spectra. This could lead to an effective imbalance in the true "target" sequences and false "decoy" sequences, since at least five of the target frames are also likely to be wrong, which in turn compromises FDR estimates.

A further possible explanation for the poor PSM sensitivity observed for the six-frame searches could be derived from the ESTScan2/EORF translations correcting errors such as frame-shifts, leading to peptides that are absent from the six-frame database. To test this, using the $FDR_{Käll}$ metric, we compared unique peptide sequences from ESTScan2 PSMs to those derived from six-frame PSMs, in Figure 2a. Only 389 peptides (~3% of ESTScan2 peptides derived from PSMs with *q*-values less than the 0.01 threshold) are unique to the ESTScan2 database searches. Most of these are expected to come from the translational corrections applied by ESTScan2. While these peptides do contribute to the higher sensitivity of the ESTScan2 search, their overall contribution is minimal. In contrast, 11 902 peptide sequences derive from PSMs found in both databases and 7917 of these have PSM *q*-values less than the 0.01 significance threshold. Hence, 3985 peptides are "lost" in the six-frame search despite having PSMs in the Mascot output with the same score but with *q*-values >0.01. Thus, the majority (91%) of the additional ESTScan2 peptides have arisen as a result of the inflated *q*-values associated with the six-frame target−decoy error modeling as opposed to correction of frame shift errors. It should also be noted that some 703 peptides are exclusive to the six-frame search as ESTScan2 did

dx.doi.org/10.1021/pr300411q | *J. Proteome Res.* 2012, 11, 5221−5234

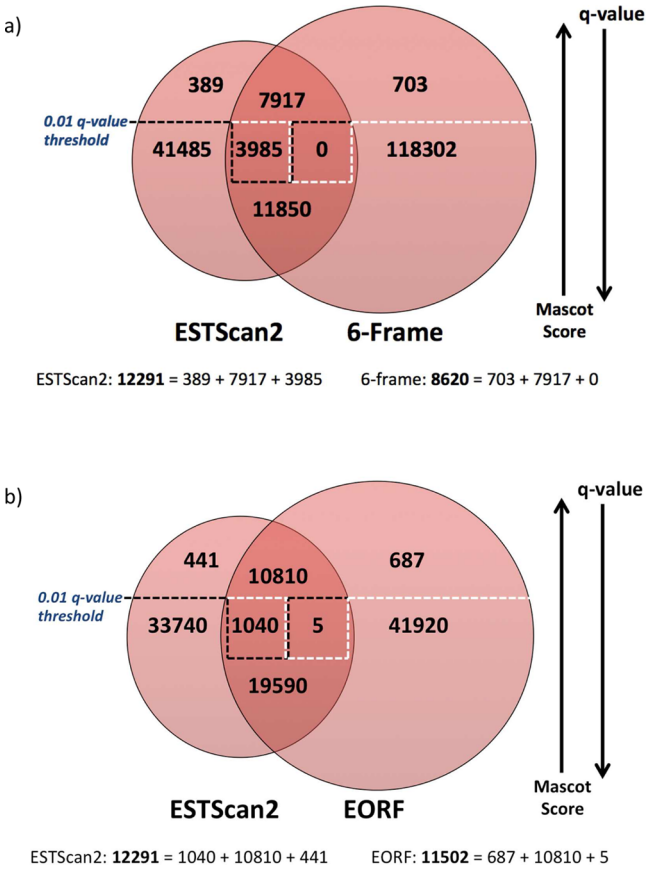


Figure 2. Overlap of peptides identified in pairwise database searches Overlap of unique peptide sequences derived from PSMs in the searches against: (a) the ESTScan2 and six-frame databases, (b) ESTScan2 and EORF databases. In both cases, FDR_{Kall} *q*-value cut-offs of 0.01 for the various searches are indicated by dotted lines, black for ESTScan2 and white for the six-frame or EORF searches. PSMs are sorted by Mascot score from low scores (bottom) to high scores (top). The majority of the unique accepted peptides identified in ESTScan2 but missed by the six-frame database were present on both databases but have *q*-values that exceed the threshold six-frame q-value threshold.

not predict a coding sequence, and thus could not be matched by any spectra. These general trends are matched when considering data at the PSM instead of the peptide level (see Supporting Information Figure S1).

In contrast, fewer peptide identifications are "lost" when comparing ESTScan2 and EORF search results, shown in Figure 2b. In this case, ESTScan2 generates more accepted PSMs and attendant unique peptides, but only 1040 additional peptides are "lost" to EORF owing to different FDR modeling while 10 810 are shared. The same trend is observed at the PSM level, where only 2% of accepted ESTScan2 PSMs were missed by EORF due to FDR differences, and 87% of accepted ESTScan2 PSMs were also accepted by EORF (see Supporting Information Figure S1).

We also note that combining the EORF and ESTScan2 databases leads to fewer PSMs at an equivalent q-value/PEP threshold, as shown in Table 1. This could be due to increased size as well as redundancy when combining similar databases (i.e., EORF and ESTScan2) where target protein sequences are more likely to be shared. Redundancy of target sequences is already known to degrade search quality; this can be addressed by merging highly similar sequences into a single entry and appending sequence variants in a compatible fashion for MS database searches.[22] One potential downside of this approach, however, is the potential loss of subtle variations that represent different biological entities that could be important for proteogenomics.

To circumnavigate this redundancy issue and integrate PSMs from separate ESTScan2 and EORF searches, we used the *FDRScore* method[52] that was initially designed to merge results from multiple search engines.[52,54,55] This approach removes redundancy at the PSM level by combining $FDR_{Käll}$ *q*-values from matched PSMs across two database searches and models database-specific PSMs independently, to derive an integrated FDR estimate. This resulted in 1522 additional PSMs compared to ESTScan2 alone and 5193 compared to the six-frame database. For our spectral data set, this results in the largest number of significant PSMs from all the approaches considered at the same nominal significance threshold of 0.01. However, this is only applicable to the *FDRScore* method when used to calculate *q*-values for concatenated target−decoy databases.

The relationship between *q*-value/PEP and Mascot score was further investigated in order to explain the poor sensitivity of the six-frame translation searches. The PEP, PEP-derived *q*-values, and empirical *q*-values ($FDR_{Käll}$) were plotted against Mascot scores for each database search (Figure 3). Figure3a confirms that the Mascot ion score for equivalent PSMs is independent of the database searched, a relationship that holds for all the paired searches we ran and supports comparisons across the different database searches. Figure 3 confirms that the PEP is clearly the more conservative approach, consistent with previous studies[36,44] and that, independent of the significance measure used, the six-frame translation PSMs have higher *q*-values/PEPs compared to the other database searches.

Biases in the target−decoy database construction methods for FDR calculations have been noted before, which can lead to effective decoy database sizes larger than the target.[44] We contend that standard six-frame translation databases may also suffer from related problems leading to inflated *q*-values and PEPs, as observed in Figure 3. The extent of this inflation can be gauged by comparing the maximum *q*-values of the different database searches. For example, Figure 3d shows that the maximum PEP-derived *q*-value (the *y*-intercept) for the six-frame translation PSMs is much greater than the equivalent *q*-values for the EORF and ESTScan2 PSMs: approximately 0.8 compared to 0.53. Similarly, the maximum $FDR_{Käll}$ *q*-value for the six-frame database search is much larger than the equivalent *q*-values from other searches (Figure 3b), and the same trend is also observed for the PEP (Figure 3c). The differences in these profiles show the effect that the choice of error model and database has on search results for high-throughput proteomics.

Collectively, these results point to the statistical modeling and not the search engine scoring that lead to a sensitivity reduction in accepted PSMs at a fixed statistical threshold for six-frame searches, and highlight the importance of redundancy removal and careful database design for estimating PSM statistics in proteogenomics. However, although we observe a broad range of FDR and PEP estimates for the same PSMs, we still do not know which ones are closest to the truth, although one might presume the six-frame values are overly conservative since this is the least sensitive search. We try to address this in the next section.

dx.doi.org/10.1021/pr300411q | *J. Proteome Res.* 2012, 11, 5221−5234

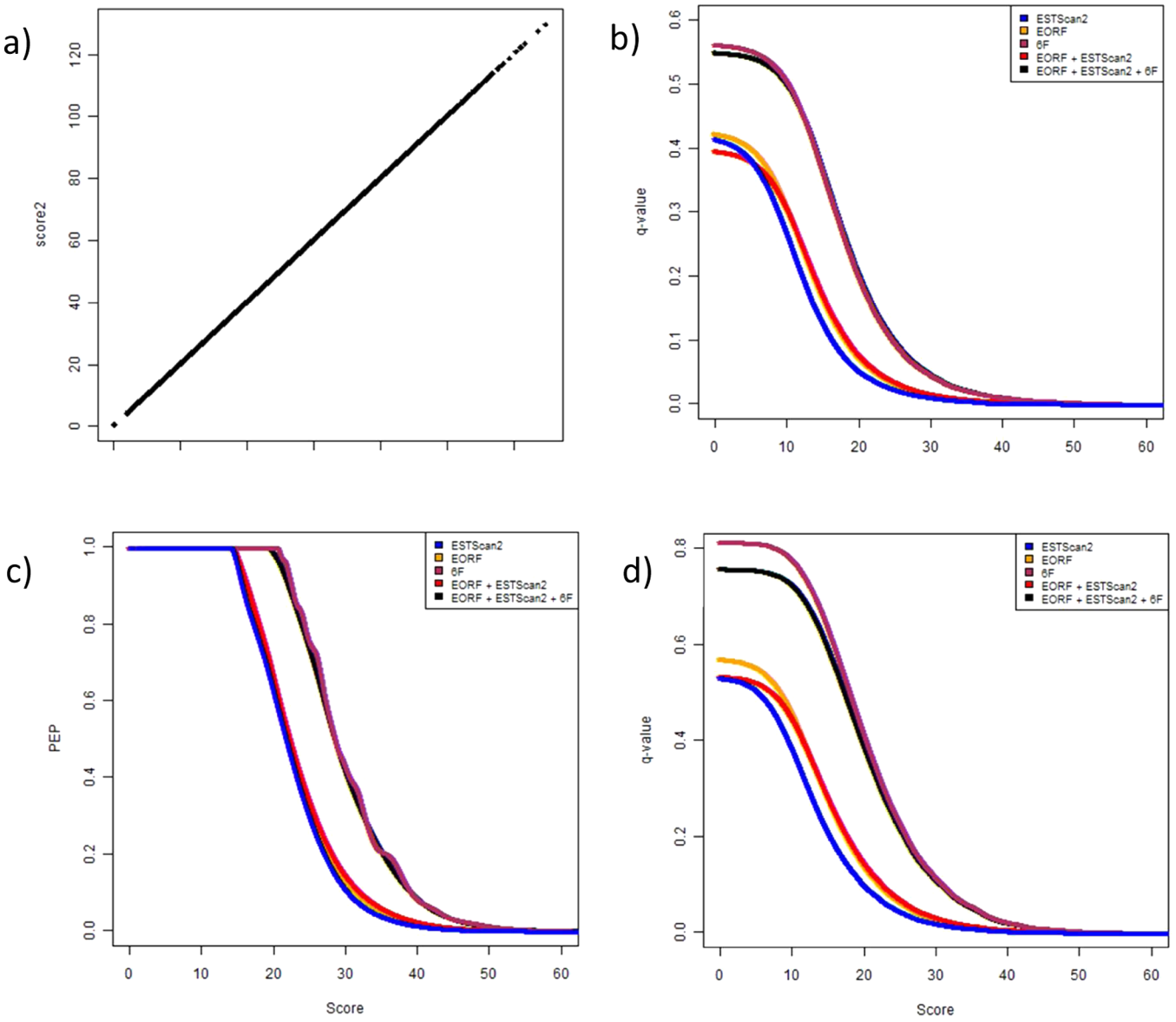


Figure 3. Variation of search statistics with Mascot score. Plots show the calculated *q*-values and PEPs for PSMs from different proteogenomic database searches and their dependence on Mascot ion score. (a) Mascot Scores of equivalent PSMs from two independent database searches are plotted, in this case ESTScan vs six-frame, although identical plots were obtained for all pairwise comparisons. (b) The *q*-values calculated using FDR_{Kall} are plotted against Mascot ion score, (c) PEPs calculated using Qvality, and (d) *q*-values calculated from Qvality, for different database search combinations. In the key, 6F denotes the six-frame searches.

The *q*-Value/PEP Values for the Six-Frame Database Search Are Overestimates Compared to Expectation

To further study the most appropriate database and error model for the six-frame search results, we investigated an alternative measure of truth for PSMs passing a given statistical threshold. We reasoned that for all EST contigs with significant BLASTX hits to the Chicken Ensembl56 set of proteins, only one of the six translated frames is "correct" based on the top-scoring BLASTX hit to Ensembl. Corresponding PSMs to this "correct" frame were also classified as "correct", all others as "incorrect".

We therefore expect that close to 100% of all PSMs are "correct" at the lowest *q*-values/PEPs and, correspondingly, target PSMs with very high *q*-values/PEPs should be false. Accordingly, PSMs with *q*-values and PEPs close to zero should be in the "correct" reading frame almost 100% of the time, whereas manifestly incorrect PSMs should be in the "correct" reading frame approximately 17% (one-sixth) of the time. Figure 4 shows that the proportion of correct reading frame PSMs is in fact close to 92.5% at the lowest *q*-values/PEPs. This small discrepancy can be explained by the fact that the reading frame with the most significant BLASTX hit does not always correspond with (and contain) all the true protein sequence. This will be mostly due to sequencing and mis-assembly errors from the ESTs generating "frameshift" errors that push the true coding sequence in to multiple frames. Therefore, we likely underestimate the proportion of correct PSMs by approximately 7% at the lowest (most significant) *q*-value/PEP thresholds due to these instances. This represents the maximum error for the estimated percentage of true positive PSMs. Figure 4a shows the percentage of correct PSMs from the total number of PSMs accepted for each q-value calculated using both the $FDR_{Käll}$ and FDR_{EG} *q*-values, as well

dx.doi.org/10.1021/pr300411q | *J. Proteome Res.* 2012, 11, 5221−5234

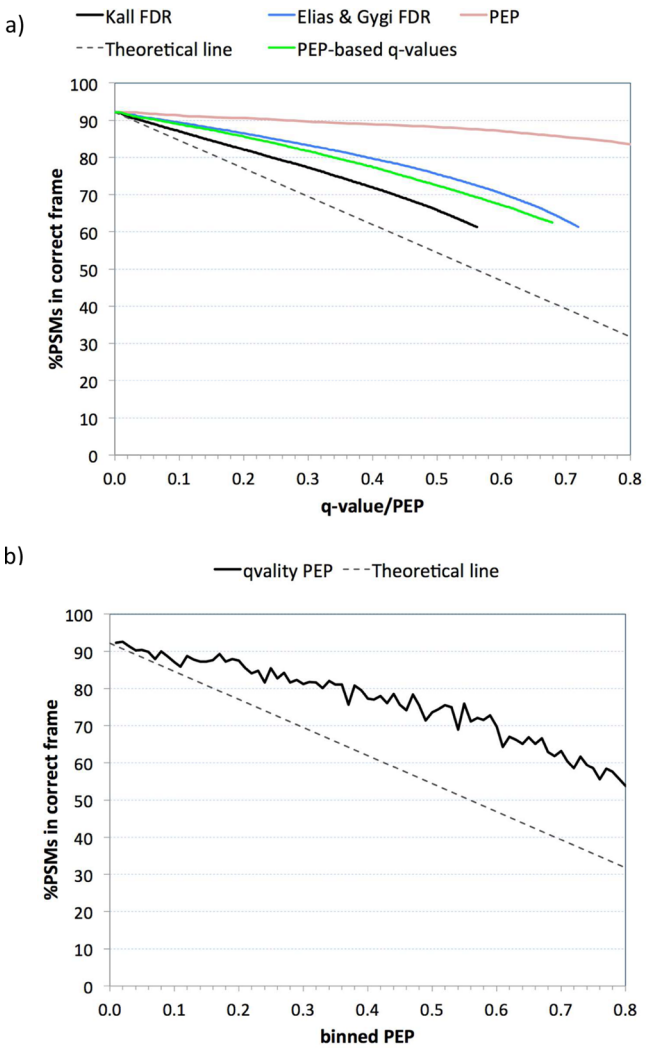


Figure 4. Estimating the proportion of true positive PSMs identified in the six-frame database search. PSMs were considered to be 'correct' if the reading frame contained the top-scoring match to an Ensembl56 protein through a BLASTX search. Plots show: (a) the percentage of 'correct' reading frame PSMs that fall below each of the three types of *q*-values and PEP, and (b) the same percentage but plotted for local qvality PEP bins of 0.01.

as from separate database searches using Qvality. The dashed line represents the percentage of PSMs falling in the "correct" reading frame that would be expected at a given *q*-value threshold, presuming the error modeling is accurate (see Supporting Information Excel File for derivation). For example, approximately 58.3% of PSMs should be in the "correct" frame at a *q*-value of 0.5, which falls slightly further still to 54% when we factor in the scaled 7% correction described above (Figure 4a). However, the observed percentages are significantly higher than would be expected by chance, at close to 66%, 73%, and 75% of PSMs assigned to the correct reading frame for the $FDR_{Käll}$, Qvality, and FDR_{EG} based *q*-values, respectively. This highlights the overly conservative nature of the *q*-value estimates for the six-frame database search. Interestingly, the $FDR_{Käll}$ *q*-value yields percentages that are closest to expectation, suggesting that for this case, at least, it is a more accurate FDR calculation.

The proportion of true positive PSMs at different PEPs was also estimated, again assigning "correct" reading frames via BLASTX searches (Figure 4b). In this case, the percentage "correct" was calculated within 0.01 bins because the PEP, unlike the *q*-value, is a local measure of significance which can become overly conservative when thresholds are applied to a list of PSMs, as discussed previously by Käll et al.[56] The extent of this conservativeness can be seen in Figure 4a, where the percentage of correct PSMs remains above 80% for PEPs above 0.8. The binned PEP data in Figure 4b is noisy compared to the cumulative data in Figure 4a, but nevertheless shows a similar trend to the *q*-value data. Again, the PEP overestimates the expected proportion of true positive PSMs and both q-value and PEP suffer from poor accuracy when deployed without correction in six-frame database searches.

Nevertheless, the PEP is a highly informative statistic in proteogenomics since a novel gene or splice-variant might only be identified via a single PSM, and we therefore need to know the likelihood of this being a correct match. Hence, it has assisted the high-throughput identification of genes[31] and is an alternative to heuristics such as the somewhat arbitrary two-peptide rule.[42] However, most tools require a decoy database in order to generate a null model, which allows tools such as Qvality to estimate PEPs more accurately.[47] Likewise, PeptideProphet utilizes decoy databases to allow PEP calculations that are free from parametric assumptions to provide accurate PEPs for different search engines and data sets.[49,57]

Six-Frame Databases Confound the Target–Decoy Assumption

The source of the atypical statistical estimates generated from the six-frame searches is not immediately obvious. Although relative differences in target and decoy databases sizes can alter the number of accepted PSMs at a given FDR threshold,[53] thereby breaking the target–decoy assumption, this does not apply here. Our target and decoy database sizes are matched in all searches, including the six-frame searches. Indeed, any difference between the number of unique target and decoy peptides can be factored into the FDR calculation using a normalization step;[40,44] this is implemented in the *Qvality* software used here for separate database searches. Regardless of the metric used to estimate significance, the fewest accepted PSMs are always observed in the six-frame database.

The unique feature of the six frame database is that only one in six target sequences is likely to contain the "true" target and most of the target database is inflated by sequences in the 'wrong frame'. This could imbalance the distributions of target and decoy PSM scores, which we investigate here.

Normally, a probability distribution of PSM scores would have a tail to the right, corresponding to the correct target PSMs.[39] Figure 5 shows the distributions of *reported* target and decoy Mascot scores for the top ranked PSMs identified in the EORF, ESTScan2 and six-frame databases for separate target and decoy database searches. It should be noted here that Mascot does not report PSMs when there is no match (presumably when there is no matching precursor ion or too few matching fragment ions). Hence, in this work, statistics are presented only on the *reported* PSMs. This has a significant effect as, for example, 369 477 target PSMs are reported from six-frame database searches compared to only 181 699 from ESTScan2, although 403 820 spectra are searched in both cases (see also Supporting Information Table S1).

The EORF and ESTSCan2 searches (Figure 5b,c) show a clear difference between the reported target and decoy score

dx.doi.org/10.1021/pr300411q | J. Proteome Res. 2012, 11, 5221–5234

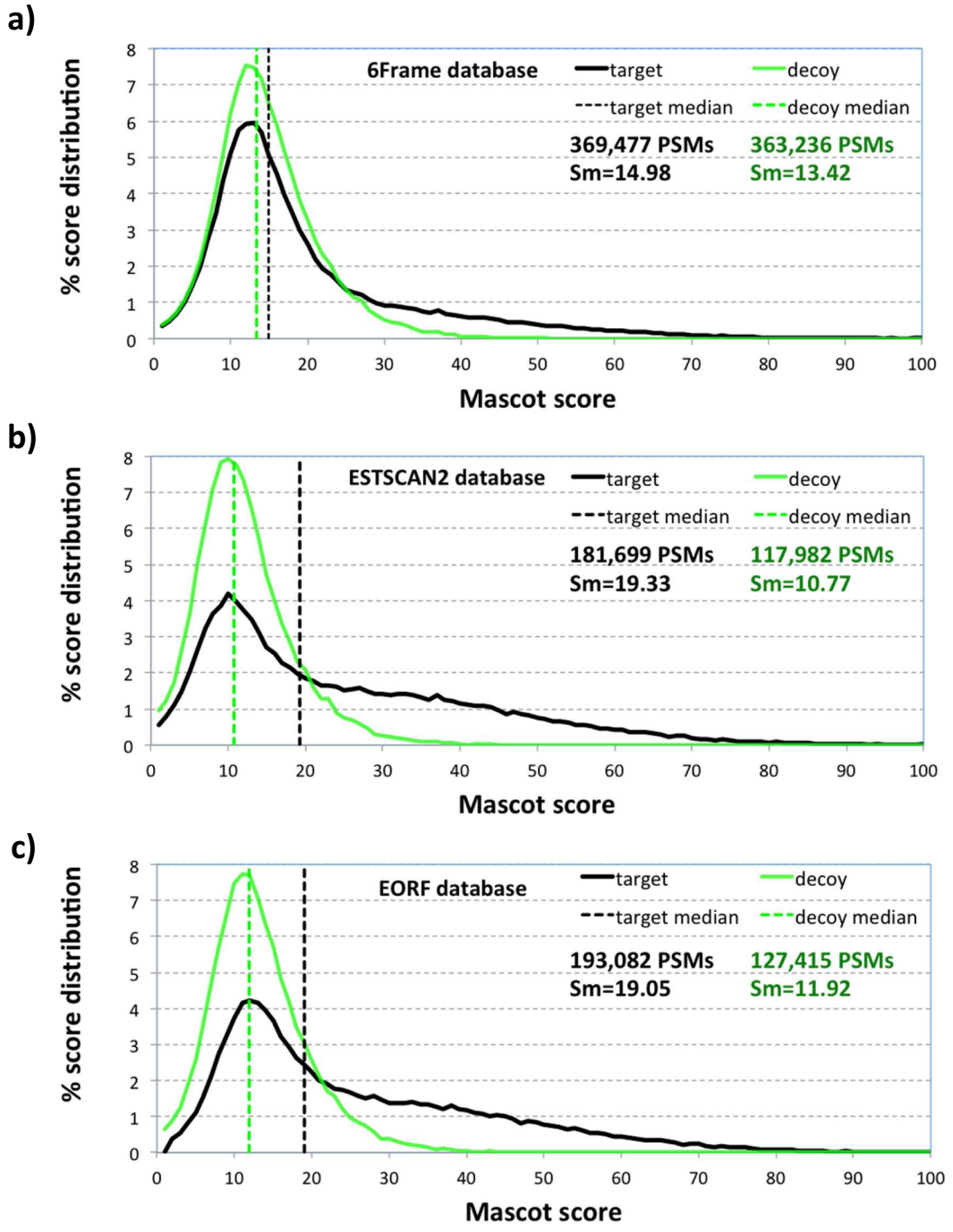


Figure 5. Mascot ion score distributions for reported target and decoy PSMs. Plots show reported target and decoy PSMs ion score distributions, for all rank 1 PSMs, when target and decoy databases were searched separately. Density plots were generated for: (a) standard six-frame database search, (b) ESTScan2 search, and (c) EORF search. The number of reported PSMs from searches of 403 820 spectra against the individual databases are also shown, demonstrating how fewer spectra are matched by Mascot for the smaller, ESTScan and EORF databases.

distributions, with the expected tail on the right of high-scoring PSMs that are likely to be correct (i.e., a low PEP or FDR). However, this clear distinction is absent for the six-frame database searches in Figure 5a, with median scores of 14.98 and 13.42 respectively for reported target and decoy PSMs. The ESTScan2 median scores for reported target and decoy PSMs are considerably different and better separated, at 19.33 and 10.77, respectively, with similar values for EORF. The increase in the median target PSM score from six-frame database to ESTScan2 appears counterintuitive since more peptides are present in the six-frame database, including the vast majority of those in the ESTScan2/EORF databases. However, as noted above, considerably fewer target PSMs are reported by Mascot for the ESTScan2 database compared to the six-frame one and these are generally of higher score (which leads to an increased median score), as is evident from Figure 5.

For the six-frame searches, the increase in the reported median decoy PSM score stems from the increased number of candidate sequences against which each spectrum can match. Indeed, Mascot reports a greater number of candidate decoy PSMs overall for the six-frame searches (363 236 compared to 117 982/127 415 in ESTScan/EORF searches). This in turn

dx.doi.org/10.1021/pr300411q | J. Proteome Res. 2012, 11, 5221−5234

would lead to more decoy PSMs out-competing their target equivalents in a concatenated database search, potentially generating more false negatives. This is indeed the case when the cumulative frequency distributions of the target and decoy PSM Mascot scores are considered for the six-frame searches in comparison to ESTScan or EORF (see Supporting Information Figure S2), where decoy PSMs are assigned higher scores in the six frame database searches for both concatenated and separate search strategies.

This inflation of the target database with 'wrong frame' sequences has consequences for FDR estimates. We illustrate this in Table 2 where we consider a search against a hypothetical standard database that produces 1000 true positive target PSMs and 10 decoy PSMs. We then consider an inflated database that contains the same target sequences plus extra sequences (i.e., alternatively translated 'wrong' frames) that are unlikely to be matched as true positives. We presume the same 1000 true PSMs will be returned from a search against this database, but false positive matches (as estimated from decoy PSMs) will likely increase (say to 20), particularly at modest search engine scores. This is because the decoy database is increased while the effective target database (i.e., "hit-able" sequences) remains the same. The resulting FDR estimate is then increased, leading to fewer accepted PSMs at the same fixed cutoff.

Table 2. Hypothetical Target and Decoy PSMs Accepted at a Fixed Score in a Standard and Inflated Database

	target PSMs	decoy PSMs	FDR at fixed score threshold
Standard database	1000 TPs		
	10 FPs	10 FPs	= 10/1000 = 0.01
Inflated database (e.g., six-frame)	1000 TPs		
	20 FPs	20 FPs	= 20/1000 = 0.02

This hypothesis is consistent with the data shown in Figure 5a where additional higher scoring decoy PSMs are identified owing to the inflated nature of the six-frame database. Conversely, the presence of the additional five 'wrong frame' target sequences increases the chances of assigning spectra to a false positive target. Indeed, roughly twice as many spectra are assigned to target PSMs by Mascot with reportable scores in the six-frame searches. Somewhat counterintuitively this lowers the median target PSM score (see Figure 5) as Mascot reports more lower scoring PSMs, presumably from hits to the five "wrong frame" sequences. Ultimately, this means that the PEP estimates are increased because they depend on the relative heights of the target and decoy distributions. Likewise, the six-frame FDR will be increased because the decoy PSMs, with relatively higher scores, would be encountered sooner when the sorted list of scores is traversed from high to low during the FDR calculation.

This is illustrated when comparing the ratio of target:decoy PSMs in a concatenated search at a fixed Mascot score between six-frame and ESTScan2 searches (see Supporting Information Figure S3). Both distributions converge toward the expected 0.50 for PSMs with low Mascot scores, although a small bias toward target PSMs remains for the six-frame database even at scores below 5. However, generally, the six-frame ratio is lower, particularly at ion scores above 15. This leads to increased numbers of decoy PSMs, which in turn leads to an increased estimate in false target PSMs and an increased FDR estimate. The same effect is observed when considering the absolute numbers of target and decoy PSMs in the six-frame database compared to a subset of it composed of just the three forward frames (see Supporting Information Figure S4). Although this database is half the size, it contains most of the true target PSMs (since most correct translations are in one of the forward frames). The numbers of reported target PSMs are almost identical between the two searches at all Mascot scores (with an average ratio of 1.05:1), while there are 1.4× more decoy PSMs on average in the six-frame search results.

The observations detailed above support the hypothesis outlined in Table 2, that the inflated nature of the six-frame translation database is the underlying reason for the over-estimation of the FDR for the six-frame searches. However, could this simply be a consequence of increased database size? Since database size will affect FDR and PEP estimates, it has been suggested that larger database sizes can reduce the variance in estimates of the number of decoys (false) PSMs, leading to better (more *precise*) estimates of significance. However, we note that in general this leads to more conservative estimates (see Supporting Information Figure S5), consistent with previous observations.[53] To confirm this, we recalculated the *q*-value for PSMs drawn from a subset of the six-frame database of identical size to ESTScan2, repeating this 1000 times to generate an average FDR profile, shown in Figure 6. This demonstrates how the atypical database composition must also be influencing the FDR calculations, since the six-frame subset *q*-value profile is more conservative than the ESTScan2 search despite the fact the database sizes are identical. The increased FDR for the six-frame searches is therefore due both to database size and atypical composition. This highlights the need to exercise caution when generating/modifying proteogenomic databases in order to provide more accurate statistical estimates.

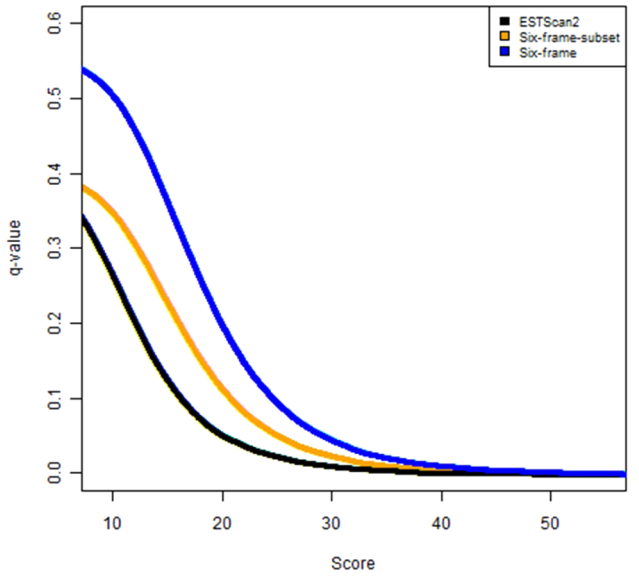


Figure 6. Effect of database size on FDR of the six-frame PSMs. Subsets of sizes equal to the ESTScan2 database were randomly sampled (1000 times) from six-frame database. The mean *q*-values were calculated from the samples to give an FDR profile with FDRs greater than the ESTScan2 PSMs, but lower than the six-frame PSMs.

dx.doi.org/10.1021/pr300411q | *J. Proteome Res.* 2012, 11, 5221−5234

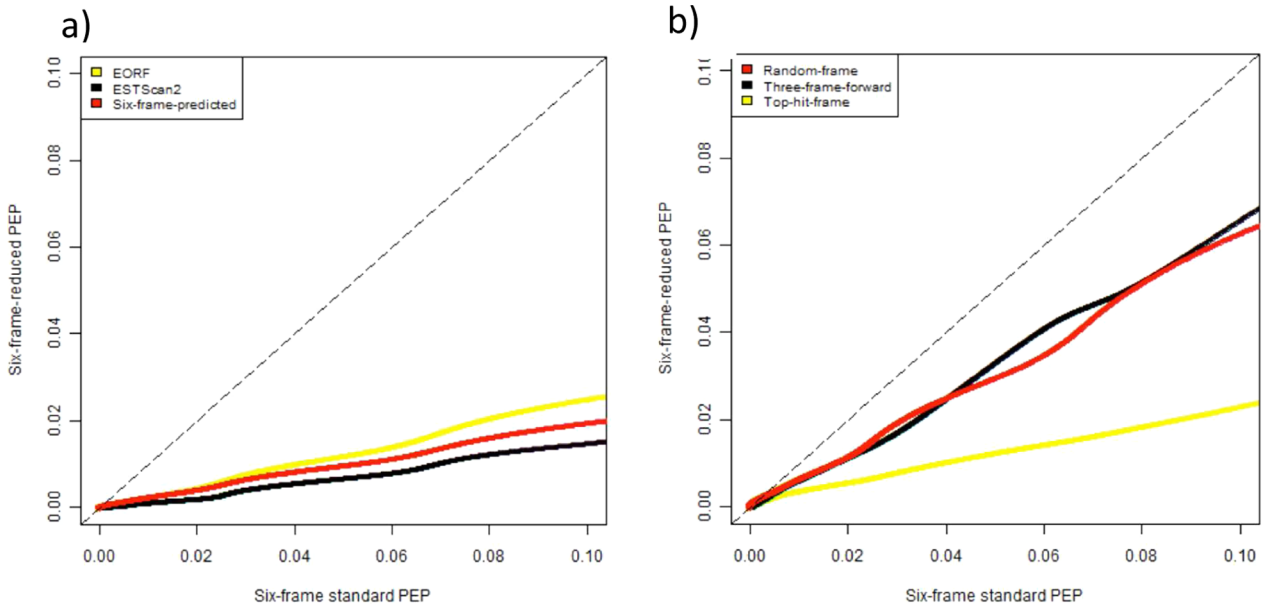


Figure 7. Comparison of equivalent PEPs from standard six-frame searches against alternate database searches. PEPs derived from several search strategies are plotted against the six-frame equivalents, with the same sequence-spectra-Mascot score. (a) PEPs derived from simple filtering approaches based on selection of a single frame by: random (*random-frame*), the most PSMs (*top-hit PSM*), or the three forward frames, are plotted against the six-frame PEP values. (b) PEPs derived from searches against the six-frame-predicted, ESTScan2 and EORF databases are plotted against the six-frame equivalents. In both plots, direct equivalence of PEP values against the standard six-frame database searches is shown as a dashed line. In all cases, selection of single frames, three forward frames, frame prediction and/or translation by EORF or ESTScan reduces the estimated PEP.

Equalizing the Target and Decoy Databases Improves the Sensitivity for the Six-Frame Searches

To estimate accurate FDRs/PEPs, the ratio of target to decoy sequences needs to be as close to 1:1 as possible, or at least properly understood and quantified. We reasoned that reducing the six-frame database to a single target and decoy sequence for each EST contig would provide a more realistic and accurate FDR estimate. A naïve but simple way to achieve this is by randomly selecting only one from each of the six target reading frames for each EST contig to generate a *random-frame* database. A second approach is to select the single frame containing the most PSMs to generate the *top-hit-frame* database. In both cases, a single target sequence is selected from the EST contigs along with its reversed (decoy) sequence. The associated PEPs were then recalculated for these modified databases and compared to the equivalent values from the standard six-frame searches, shown in Figure 7 and Table 3. As expected, this has a marked effect on PEP estimates for equivalent PSMs between the paired searches. Both methods uniformly lower the PEP compared to equivalent six-frame PSMs leading to additional unique peptide matches. Indeed, although the *random-frame* database has 83% of the original sequences removed, the number of peptides matched is 657 in excess of expectation at a PEP cutoff of 0.01, presuming that only one-sixth of the original PSMs should remain. This is an interesting finding as it shows that choosing a single frame, effectively at random for many EST contigs, improves relative sensitivity even though some "correct" frames will have been removed by chance. This would not, of course, be a viable strategy in practice since fewer PSMs overall are obtained and many genuine PSMs will be lost, but it serves to illustrate the point that less conservative statistical estimates can be achieved by filtering the database.

We tested other approaches to improve sensitivity, exploiting *a priori* knowledge on the EST contigs. Given that the direction of sequencing is usually known in transcript libraries, we observe that simply retaining just the forward three frames has an advantage (*three-frames-forward*), producing over a thousand more unique peptides with PEPs of 0.02 or less (Table 3). Similarly, considering homology to predict the mostly likely coding frame improves performance. For the EST contigs with BLASTX hits to Ensembl56, the single frame with the most significant *E*-value was retained, to generate a *six-frame predicted* database. Figure 7 shows how this approach significantly reduces the PEPs relative to the standard six-frame searches, as do EORF and ESTScan2. It should be noted though, that although the *six-frame-predicted* database led to the lowest PEP estimates, this approach yields fewer significant PSMs compared to EORF and ESTScan2, shown in Table 3.

Table 3. Unique Peptide Identifications at Different PEP Cutoffs Derived from Searches over Different Databases[a]

	PEP cutoff		
	0.01	0.02	number of target sequences
Six-frame standard	4951	5887	512,916
Random-frame	1483	1739	85,486
Three-frames-forward	5654	7050	256,458
Top-hit-frame	6624	8330	74,374
Six-frame-predicted	5601	7251	20,670
ESTScan2	7325	8907	62,161
EORF	7146	8782	67,125

[a] *Six-frame standard* refers to the standard, unfiltered six-frame translation databases. *Random-frame* refers to the subset of the six-frame database where a single frame is selected randomly for each contig. *Three-frames-forward* refers to just the forward frames only. The *six-frame-predicted* database is produced by retaining only the single frame from six for each contig that has the most significant BLASTX hit to Ensembl proteins. *Top-hit-frame* refers to the subset of the six-frame database where the frame with the most PSMs is selected for each contig.

dx.doi.org/10.1021/pr300411q | *J. Proteome Res.* 2012, 11, 5221−5234

This highlights a downside to the BLASTX-filtering approach in this instance; only about 22% of the EST contigs have significant BLAST hits whereas EORF and ESTScan2 predict protein sequences for the majority of them. The additional peptide matches from EORF and ESTScan2 searches may come from novel genes or isoforms that are not yet annotated in the Ensembl56 known gene set.

Collectively, these results demonstrate a variety of approaches to apply when considering proteogenomic searches against nucleotide databases such as EST/cDNA or genomic six-frame translations. Minimally, selecting one single candidate reading frame from the six possible leads to superior FDR and PEP estimates, and using empirical evidence to select one of the six frames via BLASTX searches against existing protein databases is better still. Nucleotide sequences that have no significant BLASTX hits still need to be dealt with, but simple strategies here can also be applied. For example, even weak BLASTX hits can be used to suggest the most likely frame that contains some level of coding features. Similarly, statistics such as codon usage can suggest "protein-like" features that point to the mostly likely frame. We note here that the "correct frame" sequences share similar amino acid composition statistics with Ensembl proteins (see Supporting Information Figure S6, Table S1). Other good practice is evident in the literature, such as the prot4EST pipeline which includes a rule-based approach that identifies the longest ORF from the six-frame translations.[58] Similarly, Adamidi et al.[12] selected the three longest ORFs from RNA-seq transcripts for database searching. Another approach, which does not require sequence homology, utilizes information about the experimental protocol used for cloning and sequencing the mRNAs. One small-scale proteomics study identified 51 novel seminal fluid proteins by searching 2D-LC-MS data against only the forward frames of translated UniGenes.[59] When applied to shotgun proteomics data this improves sensitivity relative to standard six-frame searches, a trend we also note here. Our EST data is generated from directional cloning and subsequent 5′-end sequencing of chicken cDNAs, leading to a bias where 97% of the "correct" frames from BLASTX searching are in frames one, two, or three (see Supporting Information Figure S7). Searching against this three-frame database leads to over 1000 additional peptide identifications filtered at a PEP < 0.02 (Figure 7, Table 3). This improved performance is likely the result of the reduction in erroneous targets, which is reduced to two out of three in the 3-frame search from five out of six. Although the performance is inferior to the BLAST-filtered single-frame and ESTScan/EORF searches, the three-frame approach can be used for all EST contigs and not just those with BLASTX matches. Moreover, this approach can be easily implemented for newly sequenced genomes when there are no homologues annotated within the clade, something that would create problems for EORF/ESTScan2 as these tools require codon usage information.

Ultimately, however, our most sensitive searches were derived from the EST translation tools, ESTScan2 and EORF, which generate a single translated sequence across the most likely reading frames. The most effective search strategy overall in terms of sensitivity involved the two database searches using the *FDRScore* algorithm for combining multiple search results.[52]

CONCLUSION

Our study has highlighted some of the pitfalls when searching against nucleotide databases via six-frame translations and how inflating the target database with incorrect sequences perturbs FDR and PEP estimates. Although some methods exist for reducing the peptide search space in proteogenomic studies,[9,10] most do not formally consider the inherent biases that can arise through the improper use of the target−decoy approach. We show here that naïve six-frame searching leads to over-conservative statistics and potential loss of high-quality peptide evidence for genomic annotation. This parallels any shotgun experiment that searches against very large protein databases (such as all of NCBI or UniProt), containing many "unhittable" sequences from other species, which can also lead to overestimated statistics and reduced sensitivity.

Although formally our results have been generated only from assembled ESTs, we see no reason to suppose that the general principles we observe should not apply to RNA-seq data or similar nucleotide sequences such as from gene prediction software or even raw genomic sequence. We recommend proteogenomic practitioners consider the following redundancy removal guidelines when searching against nucleotide databases.

- *Selection of most likely frame based on PSMs*. Retention of the single frame (and its decoy) with most PSMs prior to calculation of FDR/PEP provides less conservative statistics, although multipass approaches need to be performed appropriately.[41,60,61]
- *Selection of most likely frame based on homology*. BLASTX searches against a protein database, to identify the single frame mostly likely to be coding removes redundancy and improves sensitivity. If close homologues are not available, even weak matches to known proteins should enrich for coding sequence.
- *Selection of frame based on coding potential*. When no homologues are available, similarity in amino acid composition or codon usage can help select the most likely coding frame.
- *Translation software*. Gene prediction (e.g., Augustus) or EST translation software (such as ESTScan or EORF) can overcome introns and frameshifts, as well as remove redundancy. EST translators can be combined with BLASTX data to improve accuracy.[58]

Ideally, we suggest that gene prediction or EST translation software are likely to be most effective, since six possible frames are reduced to one while frameshift mutations and introns can be overcome. Moreover, presence of a transcript is *bone fide* evidence that a genomic region is at least transcribed and in most cDNA libraries the sequence is normally largely free of introns and the clonal direction is also known, so only three forward frames need be considered.

One should not lose sight of the *raison d'être* of proteogenomics, to identify novel genes and novel gene structure that may not exist in extant genome annotations or transcriptome data sets. This necessitates searching against translated genome sequence, and in such cases, use of a gene prediction tool such as Augustus,[62,63] perhaps used with reduced stringency to capture some atypical genes/gene structure, is preferable to raw searches against six-frames. Alternately, if searching against raw genomic sequence we suggest selecting one candidate translation (out of six) with the closest homology with Ensembl or UniRef proteins (i.e., similar

to the six-frame-predicted database), in addition to applying compositional filters to limit regions unlikely to be coding. This should improve the overall sensitivity of the searches while improving the statistical modeling of incorrect PSMs.

It is also worth noting an additional downside to selecting a single reading frame, when several true coding regions overlap at a single genomic locus but only a single one is selected *in silico* (the INK4/ARF locus is one such example). However, this might be overcome by screening the six-frame database for contigs that contain multiple high-confidence PSMs in different reading frames. Equally, transcriptomic data might very well capture the two independent but overlapping coding regions in different clusters.

Finally, it should be stressed that care needs to be taken when applying target−decoy strategies to shotgun proteomics data and proteogenomics data in particular. As several authors have pointed out, if the underlying database or search engine is not compliant with the assumptions of the target−decoy approach, significance values can be underestimated.[41,43,60,61] This is particularly acute for multipass search strategies that change the target−decoy structure between search phases. Here, we ensure the numbers of candidate target and decoy sequences in the database are the same, though we do favor target sequences with certain properties (i.e., those likely to be coding, have the most PSMs, etc.), which could introduce some biases. Nevertheless, the evidence that standard six-frame searches considerably overestimate FDRs/PEPs seems overwhelming, which highlights the need for databases to be checked for compliance with the target−decoy approach (in addition to the search engines). It is clear that more work is needed to produce better error models and databases to develop more reliable statistical confidence estimates, which would greatly benefit proteogenomics by helping to achieve a more comprehensive representation of the proteome.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Figures S1−S7 and Table S1, Venn diagrams of overlap between search results at the PSM level, cumulative frequency plots of mascot ion scores, percentage of target PSMs versus mascot ion score, comparisons of six-frame and three-frame search statistics, scatterplots of accepted peptides versus database size, amino acid composition histograms, a table of distances between amino acid frequency vectors for "correct" and "other" reading frames in the six frame searches, and BLASTX frame search statistics for the EST contigs with significant matches in Ensembl chicken proteome. Excel spreadsheet containing the calculations used to estimate the percentage of correct frame PSMs in the six-frame database at different false discovery rates. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: simon.hubbard@manchester.ac.uk. Tel: +44 1613068930.

Present Address

§The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, U.K.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Julian Selley and Craig Lawless for their kind assistance with database provision for Mascot searches, Andrew Jones at the University of Liverpool for useful comments on the manuscript, and particularly Karthryn Lilley at the University of Cambridge for supplying the chicken DT40 mass spectrometry data. We thank the BBSRC for funding in the form of a studentship to P.B. and via grant funding to S.J.H. (BB/I000631/1). I.M.O. is supported by a Royal Society of Edinburgh Scottish Government Fellowship co-funded by Marie Curie Actions and the UK Medical Research Council (MRC).

■ REFERENCES

(1) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhaus, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. Systems-wide perturbation analysis with near complete coverage of the yeast proteome by single-shot UHPLC runs on a bench-top Orbitrap. *Mol. Cell. Proteomics* **2012**, *11* (3), No. M111.013722.

(2) Schrimpf, S. P.; Weiss, M.; Reiter, L.; Ahrens, C. H.; Jovanovic, M.; Malmstroem, J.; Brunner, E.; Mohanty, S.; Lercher, M. J.; Hunziker, P. E.; Aebersold, R.; von Mering, C.; Hengartner, M. O. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol.* **2009**, *7* (3), No. e48.

(3) Chaerkady, R.; Kelkar, D. S.; Muthusamy, B.; Kandasamy, K.; Dwivedi, S. B.; Sahasrabuddhe, N. A.; Kim, M. S.; Renuse, S.; Pinto, S. M.; Sharma, R.; Pawar, H.; Sekhar, N. R.; Mohanty, A. K.; Getnet, D.; Yang, Y.; Zhong, J.; Dash, A. P.; Maccallum, R. M.; Delanghe, B.; Mlambo, G.; Kumar, A.; Keshava Prasad, T. S.; Okulate, M.; Kumar, N.; Pandey, A. A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. *Genome Res.* **2011**, *21* (11), 1872−1881.

(4) Castellana, N. E.; Payne, S. H.; Shen, Z. X.; Stanke, M.; Bafna, V.; Briggs, S. P. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (52), 21034−21038.

(5) Merrihew, G. E.; Davis, C.; Ewing, B.; Williams, G.; Kall, L.; Frewen, B. E.; Noble, W. S.; Green, P.; Thomas, J. H.; MacCoss, M. J. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.* **2008**, *18* (10), 1660−1669.

(6) Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **2008**, *320* (5878), 938−941.

(7) Wang, X.; Slebos, R. J.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009−1017.

(8) Ching, A. T.; Paes Leme, A. F.; Zelanis, A.; Rocha, M. M.; Furtado, M. D.; Silva, D. A.; Trugilho, M. R.; Rocha, S. L.; Perales, J.; Ho, P. L.; Serrano, S. M.; Junqueira-de-Azevedo, I. L. Venomics profiling of *Thamnodynastes strigatus* unveils matrix metalloproteinases and other novel proteins recruited to the toxin arsenal of rear fanged snakes. *J. Proteome Res.* **2012**, *11* (2), 1152−1162.

(9) Edwards, N. J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.* **2007**, *3*, 102.

(10) Robinson, M. W.; Menon, R.; Donnelly, S. M.; Dalton, J. P.; Ranganathan, S. An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen Fasciola hepatica: proteins associated with invasion and infection of the mammalian host. *Mol. Cell. Proteomics* **2009**, *8* (8), 1891−1907.

(11) May, P.; Wienkoop, S.; Kempa, S.; Usadel, B.; Christian, N.; Rupprecht, J.; Weiss, J.; Recuenco-Munoz, L.; Ebenhoh, O.; Weckwerth, W.; Walther, D. Metabolomics- and proteomics-assisted

genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics* **2008**, *179* (1), 157−166.
(12) Adamidi, C.; Wang, Y.; Gruen, D.; Mastrobuoni, G.; You, X.; Tolle, D.; Dodt, M.; Mackowiak, S. D.; Gogol-Doering, A.; Oenal, P.; Rybak, A.; Ross, E.; Sanchez Alvarado, A.; Kempa, S.; Dieterich, C.; Rajewsky, N.; Chen, W. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res.* **2011**, *21* (7), 1193−1200.
(13) Brosch, M.; Saunders, G. I.; Frankish, A.; Collins, M. O.; Yu, L.; Wright, J.; Verstraten, R.; Adams, D. J.; Harrow, J.; Choudhary, J. S.; Hubbard, T. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and ″resurrected″ pseudogenes in the mouse genome. *Genome Res.* **2011**, *21* (5), 756−767.
(14) Tanner, S.; Shen, Z. X.; Ng, J.; Florea, L.; Guigo, R.; Briggs, S. P.; Bafna, V. Improving gene annotation using peptide mass spectrometry. *Genome Res.* **2007**, *17* (2), 231−239.
(15) Kalume, D. E.; Peri, S.; Reddy, R.; Zhong, J.; Okulate, M.; Kumar, N.; Pandey, A. Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics* **2005**, *6*, 128.
(16) de Souza, G. A.; Arntzen, M. O.; Fortuin, S.; Schurch, A. C.; Malen, H.; McEvoy, C. R.; van Soolingen, D.; Thiede, B.; Warren, R. M.; Wiker, H. G. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol. Cell. Proteomics* **2011**, *10* (1), No. M110.002527.
(17) Baudet, M.; Ortet, P.; Gaillard, J. C.; Fernandez, B.; Guerin, P.; Enjalbal, C.; Subra, G.; de Groot, A.; Barakat, M.; Dedieu, A.; Armengaud, J. Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol. Cell. Proteomics* **2010**, *9* (2), 415−426.
(18) Blakeley, P.; Siepen, J. A.; Lawless, C.; Hubbard, S. J. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* **2010**, *10* (6), 1127−40.
(19) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H. K.; Lin, B. Y.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L. H.; Yi, E. C.; Zhang, H.; Aebersold, R. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005**, *6* (1), No. R9.
(20) Findlay, G. D.; MacCoss, M. J.; Swanson, W. J. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in Drosophila. *Genome Res.* **2009**, *19* (5), 886−896.
(21) Prasad, T. S.; Harsha, H. C.; Keerthikumar, S.; Sekhar, N. R.; Selvan, L. D.; Kumar, P.; Pinto, S. M.; Muthusamy, B.; Subbannayya, Y.; Renuse, S.; Chaerkady, R.; Mathur, P. P.; Ravikumar, R.; Pandey, A. Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. *J. Proteome Res.* **2012**, *11* (1), 247−260.
(22) de Souza, G. A.; Arntzen, M. O.; Wiker, H. G. MSMSpdbb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. *Bioinformatics* **2010**, *26* (5), 698−699.
(23) Iseli, C.; Jongeneel, C. V.; Bucher, P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 138−148.
(24) Fukunishi, Y.; Hayashizaki, Y. Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics* **2001**, *5* (2), 81−87.
(25) Gouzy, J.; Carrere, S.; Schiex, T. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* **2009**, *25* (5), 670−671.
(26) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(27) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of post translationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626−4639.
(28) Craig, R.; Beavis, R. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(29) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(30) Sevinsky, J. R.; Cargile, B. J.; Bunger, M. K.; Meng, F.; Yates, N. A.; Hendrickson, R. C.; Stephenson, J. L. Whole genome searching with shotgun proteomic data: Applications for genome annotation. *J. Proteome Res.* **2008**, *7* (1), 80−88.
(31) Borchert, N.; Dieterich, C.; Krug, K.; Schutz, W.; Jung, S.; Nordheim, A.; Sommer, R. J.; Macek, B. Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Res.* **2010**, *20* (6), 837−846.
(32) Bindschedler, L. V.; Burgis, T. A.; Mills, D. J. S.; Ho, J. T. C.; Cramer, R.; Spanu, P. D. In Planta Proteomics and Proteogenomics of the Biotrophic Barley Fungal Pathogen Blumeria graminis f. sp hordei. *Mol. Cell. Proteomics* **2009**, *8* (10), 2368−2381.
(33) Brunner, E.; Ahrens, C. H.; Mohanty, S.; Baetschmann, H.; Loevenich, S.; Potthast, F.; Deutsch, E. W.; Panse, C.; de Lichtenberg, U.; Rinner, O.; Lee, H.; Pedrioli, P. G. A.; Malmstrom, J.; Koehler, K.; Schrimpf, S.; Krijgsveld, J.; Kregenow, F.; Heck, A. J. R.; Hafen, E.; Schlapbach, R.; Aebersold, R. A high-quality catalog of the Drosophila melanogaster proteome. *Nat. Biotechnol.* **2007**, *25* (5), 576−583.
(34) Gupta, N.; Benhamida, J.; Bhargava, V.; Goodman, D.; Kain, E.; Kerman, I.; Nguyen, N.; Ollikainen, N.; Rodriguez, J.; Wang, J.; Lipton, M. S.; Romine, M.; Bafna, V.; Smith, R. D.; Pevzner, P. A. Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* **2008**, *18* (7), 1133−1142.
(35) Jaffe, J. D.; Berg, H. C.; Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004**, *4* (1), 59−77.
(36) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7* (1), 40−44.
(37) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* **2005**, *5* (13), 3475−3490.
(38) Storey, J. D. A direct approach to false discovery rates. *J. R. Statist. Soc. B* **2002**, *64*, 479−498.
(39) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29−34.
(40) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.
(41) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (7), 1111−1120.
(42) Gupta, N.; Pevzner, P. A., False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* **2009**, *8* (9), 4173−4181.
(43) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092−2123.
(44) Wang, G.; Wu, W. W.; Zhang, Z.; Masilamani, S.; Shen, R. F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* **2009**, *81* (1), 146−159.
(45) Granholm, V.; Kall, L. Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **2011**, *11* (6), 1086−1093.
(46) Kall, L.; Storey, J. D.; Noble, W. S. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008**, *24* (16), i42−i48.

(47) Kall, L.; Storey, J. D.; Noble, W. S. QVALITY: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* **2009**, *25* (7), 964−966.
(48) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.
(49) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 254−265.
(50) Boardman, P. E.; Sanz-Ezquerro, J.; Overton, I. M.; Burt, D. W.; Bosch, E.; Fong, W. T.; Tickle, C.; Brown, W. R. A.; Wilson, S. A.; Hubbard, S. J. A comprehensive collection of chicken cDNAs. *Curr. Biol.* **2002**, *12* (22), 1965−1969.
(51) Hall, S. L.; Hester, S.; Griffin, J. L.; Lilley, K. S.; Jackson, A. P. The organelle proteome of the DT40 lymphocyte cell line. *Mol. Cell. Proteomics* **2009**, *8* (6), 1295−1305.
(52) Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **2009**, *9* (5), 1220−1229.
(53) Fitzgibbon, M.; Li, Q.; McIntosh, M. Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.* **2008**, *7* (1), 35−39.
(54) Kwon, T.; Choi, H.; Vogel, C.; Nesvizhskii, A. I.; Marcotte, E. M. MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. *J. Proteome Res.* **2011**, *10* (7), 2949−2958.
(55) Alves, G.; Wu, W. W.; Wang, G.; Shen, R. F.; Yu, Y. K. Enhancing peptide identification confidence by combining search methods. *J. Proteome Res.* **2008**, *7* (8), 3102−3113.
(56) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 47−50.
(57) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), No. M111.007690.
(58) Wasmuth, J. D.; Blaxter, M. L. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinf.* **2004**, *5*, 187.
(59) Walters, J. R.; Harrison, R. G. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in heliconius butterflies. *Mol. Biol. Evol.* **2010**, *27* (9), 2000−2013.
(60) Bern, M.; Kil, Y. J. Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies". *J. Proteome Res.* **2011**, *10* (4), 2123−2127.
(61) Everett, L. J.; Bierl, C.; Master, S. R. Unbiased statistical analysis for multi-stage proteomic search strategies. *Journal of proteome research* **2010**, *9* (2), 700−707.
(62) Stanke, M.; Tzvetkova, A.; Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **2006**, *7* (Suppl1), S11.1−S11.8.
(63) Brosch, M.; Saunders, G. I.; Frankish, A.; Collins, M. O.; Yu, L.; Wright, J.; Verstraten, R.; Adams, D. J.; Harrow, J.; Choudhary, J. S.; Hubbard, T. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and ″resurrected″ pseudogenes in the mouse genome. *Genome Res.* **2011**, *21* (5), 756−767.